yang_seonhyeHW19

Seonhye Yang

3/5/2019

Question 1

```
library(alr4)
```

```
## Loading required package: car
```

```
## Loading required package: carData
```

```
## Loading required package: effects
```

```
## lattice theme set by effectsTheme()
## See ?effectsTheme for details.
```

```
attach(lakemary)
```

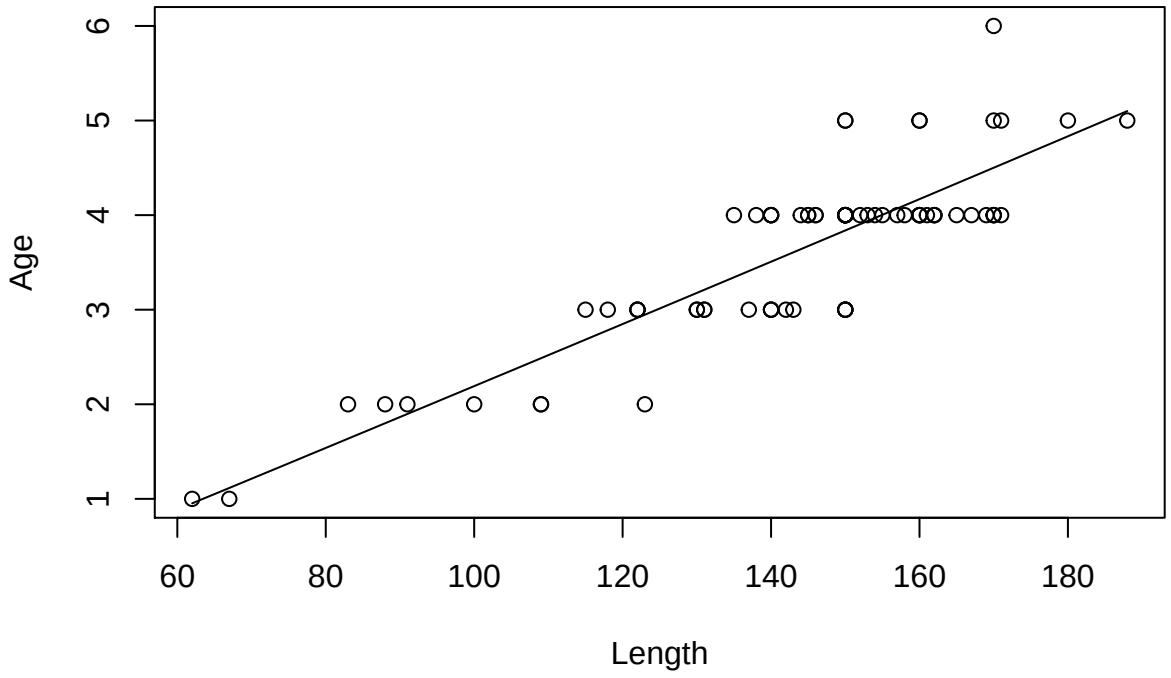
Part A

```
Length2 <- Length^2
model <- lm(Age ~ Length + Length2)
summary(model)
```

```
##
## Call:
## lm(formula = Age ~ Length + Length2)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.94688 -0.30999  0.03862  0.27529  1.49908
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -1.048e+00  1.035e+00  -1.013   0.3145
## Length       3.206e-02  1.631e-02   1.965   0.0531 .
## Length2      3.440e-06  6.319e-05   0.054   0.9567
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4838 on 75 degrees of freedom
## Multiple R-squared:  0.7349, Adjusted R-squared:  0.7278
## F-statistic:   104 on 2 and 75 DF,  p-value: < 2.2e-16
```

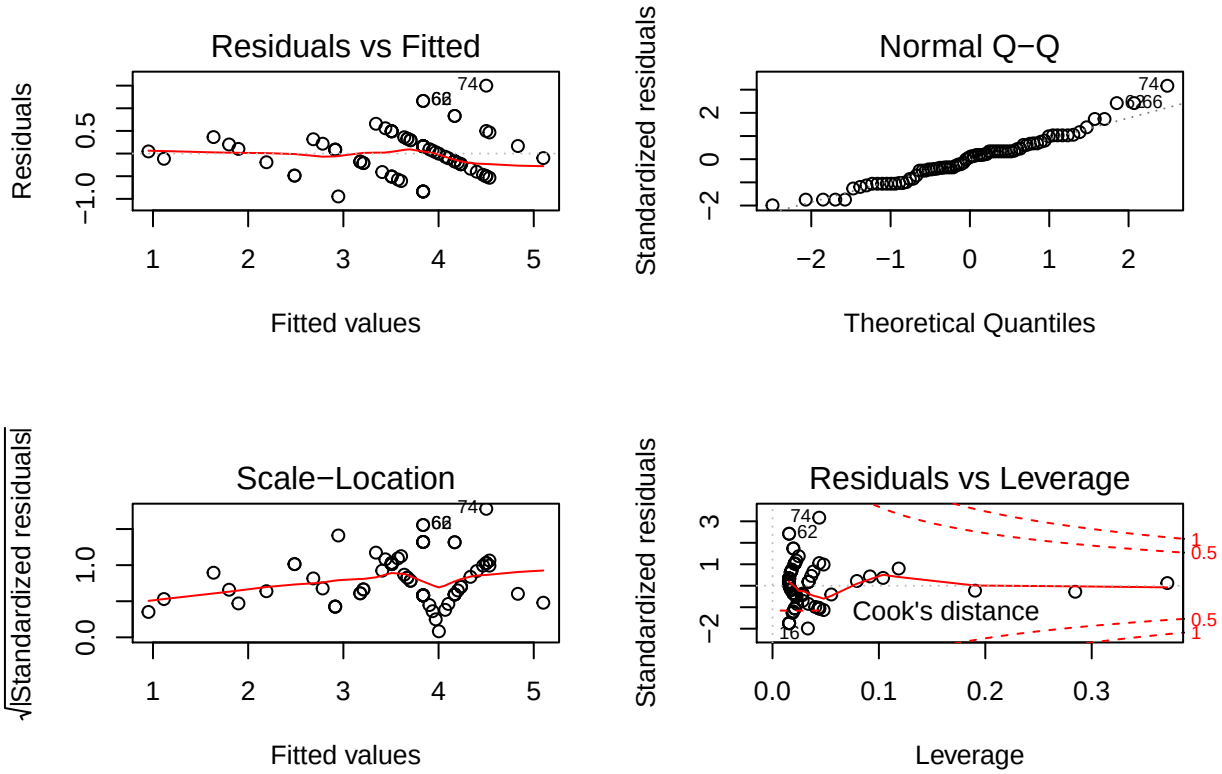
Part B

```
x.grid <- seq(min(Length), max(Length), len=100)
plot(Length, Age)
lines(x.grid, predict(model, list(Length=x.grid, Length2=x.grid^2)))
```



Part C

```
par(mfrow=c(2,2))
plot(model)
```



Looking at the Residuals VS Fitted plot, we see a flat trendline, and regularly spaced residuals, indicating a good model. The Normal Q-Q plot is also pretty much linear, indicating a good model as well. Our Scale-Location plot shows some deviation near the end, but the generally flat trendline, and generally well spaced residuals indicates a decently fit model. Finally, we see only a few points with high leverage, meaning our model represents most of our data well. Overall, these diagnostic plots indicated a decently fit model.

Question 2

Part A

```
if (!require("EnvStats")) install.packages("EnvStats")
```

```
## Loading required package: EnvStats
```

```
##
## Attaching package: 'EnvStats'
```

```
## The following object is masked from 'package:car':
##
##     qqPlot
```

```
## The following objects are masked from 'package:stats':
##
##     predict, predict.lm
```

```
## The following object is masked from 'package:base':
##
##     print.default
```

```r
library(EnvStats)
```

```r
anovaPE(model)
```

```
##                   Df Sum Sq Mean Sq  F value    Pr(>F)
## Length             1 48.662  48.662 152.0584 4.938e-15 ***
## Length2            1  0.001   0.001   0.0022    0.9631
## Lack of Fit       36  5.074   0.141   0.4404    0.9927
## Pure Error        39 12.481   0.320
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Part B

Looking at the Lack of Fit value of 5.074, and it's P value of 0.9927, we can firstly reject the null hypothesis and state that this model is a good fit for our data. In addition to this, compared to the pure error, the Lack of Fit error is nearly $\frac{1}{3}$ of the pure, indicating that most errors occuring due to the omission of important terms is gone.

Question 3

Part A

```r
library(readr)
library(data.table)
hybrid <- fread("http://users.stat.ufl.edu/~winner/data/hybrid_reg.csv")
attach(hybrid, warn.conflicts = F)
fit1=lm(msrp~mpgmpge+accelrate,data=hybrid)
fit2=lm(msrp~mpgmpge*accelrate,data=hybrid)
fit3=lm(msrp~mpgmpge*poly(accelrate,2,raw=T),data=hybrid)
summary(fit1)
```

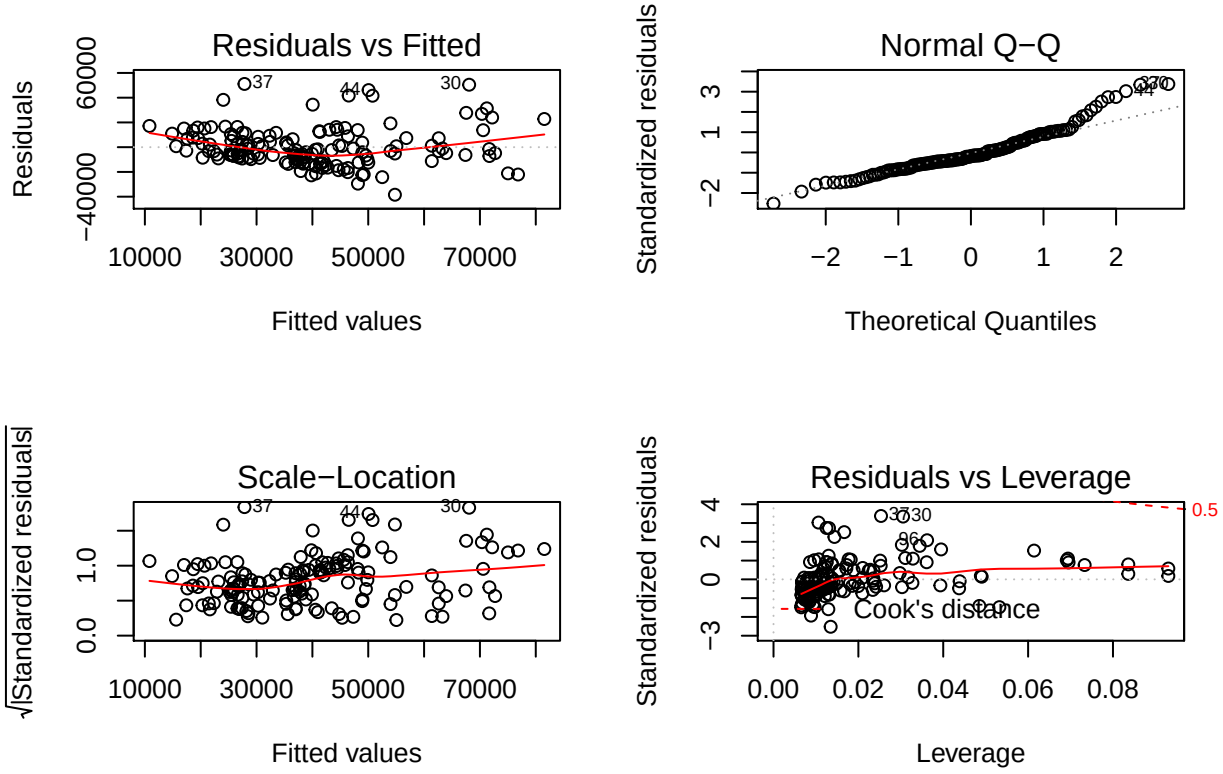
```
##
## Call:
## lm(formula = msrp ~ mpgmpge + accelrate, data = hybrid)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -38435  -8709  -2836   7755  51093
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -12309.88    7246.60  -1.699   0.0914 .
## mpgmpge       -131.48      73.85  -1.780   0.0770 .
## accelrate     4740.14     461.21  10.278   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 15330 on 150 degrees of freedom
## Multiple R-squared:  0.4945, Adjusted R-squared:  0.4878
## F-statistic: 73.37 on 2 and 150 DF,  p-value: < 2.2e-16
```

```
summary(fit2)
```

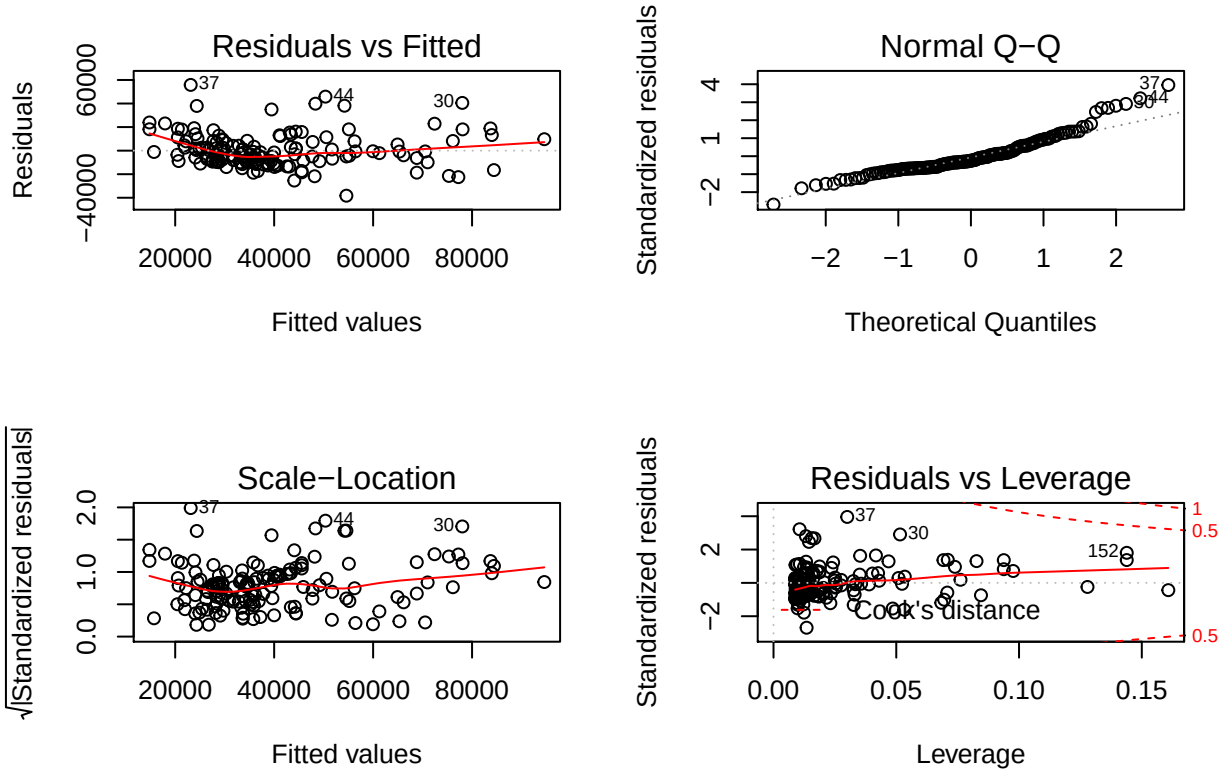
```
##
## Call:
## lm(formula = msrp ~ mpgmpge * accelrate, data = hybrid)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -38256  -9278  -3541   7374  55797
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)       -75673.13   14838.31  -5.100 1.02e-06 ***
## mpgmpge             1870.10     422.81   4.423 1.87e-05 ***
## accelrate          10440.06    1263.58   8.262 7.19e-14 ***
## mpgmpge:accelrate   -186.49      38.87  -4.798 3.85e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 14320 on 149 degrees of freedom
## Multiple R-squared:  0.5622, Adjusted R-squared:  0.5533
## F-statistic: 63.77 on 3 and 149 DF,  p-value: < 2.2e-16
```

Part B

```
par(mfrow=c(2,2))
plot(fit1)
```



```r
par(mfrow=c(2,2))
plot(fit2)
```



Beginning with the Residuals VS Fitted Plot, both models have relatively flat trendlines, however, model 2 has more lightly groups residuals, indicating a potential problem. Both have good Normal Q-Q plots, with model two being slightly more linear. Both models have linear Scale-Location trendlines, but again, model 2 seems to have come grouping. Finally, looking at Residuals vs Leverage, we can see that the leverage is overall lower in model 1, vs model 2, so the points are more consistently represented. This would mean that I would pick model 1 over model 2.

Question 4

Part A

$$\text{MSRP}_i = \beta_0 + \text{MPGMPGE}\beta_1 + \text{Accelerate}^2\beta_2 + (\text{MPGMPGE})(\text{Accelerate}^2)\beta_3$$

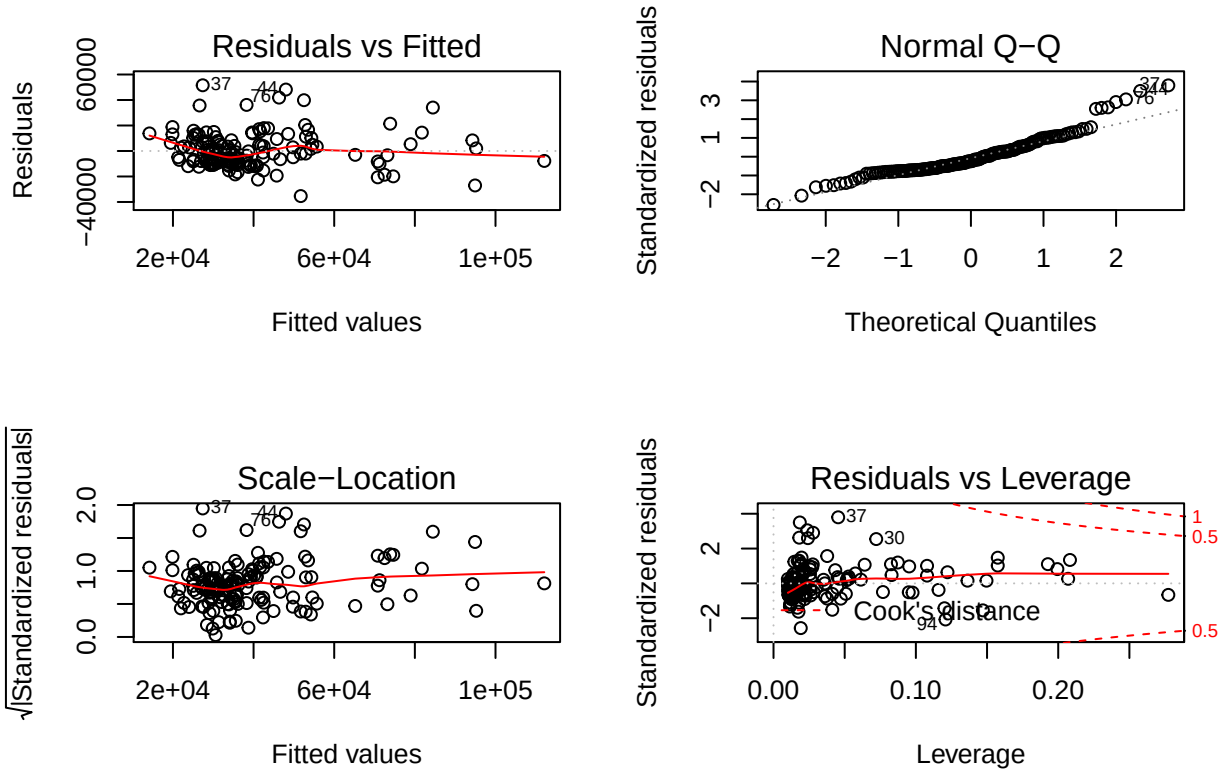
Part B

```r
summary(fit3)
```

```
##
## Call:
## lm(formula = msrp ~ mpgmpge * poly(accelrate, 2, raw = T), data = hybrid)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -35355  -9433  -3343   7287  51546
```

```
##
## Coefficients:
##                                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)                          95112.48   56325.44   1.689  0.09341
## mpgmpge                              -2993.08    1526.12  -1.961  0.05174
## poly(accelrate, 2, raw = T)1        -18388.91    9048.62  -2.032  0.04393
## poly(accelrate, 2, raw = T)2          1194.39     363.78   3.283  0.00128
## mpgmpge:poly(accelrate, 2, raw = T)1   669.75     263.23   2.544  0.01198
## mpgmpge:poly(accelrate, 2, raw = T)2   -37.06      11.45  -3.236  0.00150
##
## (Intercept)                          .
## mpgmpge                              .
## poly(accelrate, 2, raw = T)1         *
## poly(accelrate, 2, raw = T)2         **
## mpgmpge:poly(accelrate, 2, raw = T)1 *
## mpgmpge:poly(accelrate, 2, raw = T)2 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 13900 on 147 degrees of freedom
## Multiple R-squared:  0.5925, Adjusted R-squared:  0.5787
## F-statistic: 42.76 on 5 and 147 DF,  p-value: < 2.2e-16
```

```
par(mfrow=c(2,2))
plot(fit3)
```



Compared to the plots of Fit 1 and 2, Fit 3 is quite similar. Although it shows quite similar characteristics in all its plots, we can see that the spacing in both the Residuals VS Fitted and Scale-Location plots is better here than in the previous fits, in addition the Normal Q-Q line is more linear. Finally, the Residuals are more tighly grouped. This all indicates that Fit 3 is also a valid model.

Question 5

```
anova(fit1, fit2)
```

```
## Analysis of Variance Table
##
## Model 1: msrp ~ mpgmpge + accelrate
## Model 2: msrp ~ mpgmpge * accelrate
##   Res.Df        RSS Df  Sum of Sq      F   Pr(>F)
## 1    150 3.5257e+10
## 2    149 3.0538e+10  1 4718657759 23.023 3.85e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
anova(fit2, fit3)
```

```
## Analysis of Variance Table
##
## Model 1: msrp ~ mpgmpge * accelrate
## Model 2: msrp ~ mpgmpge * poly(accelrate, 2, raw = T)
##   Res.Df        RSS Df  Sum of Sq      F   Pr(>F)
## 1    149 3.0538e+10
## 2    147 2.8419e+10  2 2118947429 5.4802 0.005064 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Looking at the results, we should use Fit 3 to predict MSRP as although it may not have the smallest P value, it is well within bounds for us to reject the null hypothesis. Also, the F-statistic is far lower here, showing less dispersion, and a smaller sum of squares. All of this leads to Fit 3 being the best.